DOCUMENT RESUME

ED 413 752 FL 024 773

AUTHOR Proszeky, Gabor
TITLE Humor (High-Speed Unification Morphology): A Morphological System for Corpus Analysis.
PUB DATE 1995-00-00
NOTE 11p.; In: Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995); see FL 024 759.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computational Linguistics; *Computer Software; Computer Software Development; Discourse Analysis; Foreign Countries; *Language Processing; *Language Research; Linguistic Theory; *Morphology (Languages); *Structural Analysis (Linguistics); Word Processing
IDENTIFIERS Europe; Europe (East); *Language Corpora

ABSTRACT

Humor, a reversible, string-based unification approach for lemmatizing and disambiguating language data, has been used for both language corpus analysis and creation of a variety of linguistic software applications such as spell-checking. The system is language-independent, allowing multilingual applications for a variety of language types. Its Hungarian version, the largest and most precise implementation, contains nearly 100,000 stems. The system has been tested rigorously by both linguists and end-users of word-processing tools. Humor-based linguistic modules have been licensed by major software producers, and the lemmatizer has been used in lexicographic research since 1991. One tool provides disambiguation, tagging, and parsing functions. The system can describe various natural languages, including both Eastern European and non-Eastern European languages. Several Humor subsystems for different purposes (lemmatizing, hyphenating, spell-checking/correcting, grammar checking) are commercially available, and have been built into several major word-processing and full-text retrieval systems. An inflectional thesaurus and a series of intelligent bilingual dictionaries have also been developed. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 413 752

Humor
(High-speed Unification Morphology)
A Morphological System for Corpus Analysis

Gábor Prószéky

MorphoLogic
Németvölgyi út 25.
H-1126 Budapest, Hungary
Tel.: +361 2122 429
Fax: +361 2018 355
E-mail: h6109pro@ella.hu

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Norbert
Volz
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

FL024773

BEST COPY AVAILABLE

1. The Humor morphological system

Humor, a reversible, string-based, unification approach for lemmatizing and disambiguation has been used for both corpus analysis in the Research Institute for Linguistics and creating a variety of other lingware applications, like spell-checking, hyphenation etc. for the wide public. The system is language independent, that is, it allows multilingual applications; besides agglutinative languages (e.g., Hungarian, Turkish) and highly inflectional languages (e.g., Polish, Rumanian), it has been applied to languages of major economic and demographic significance (e.g., English, German, French).

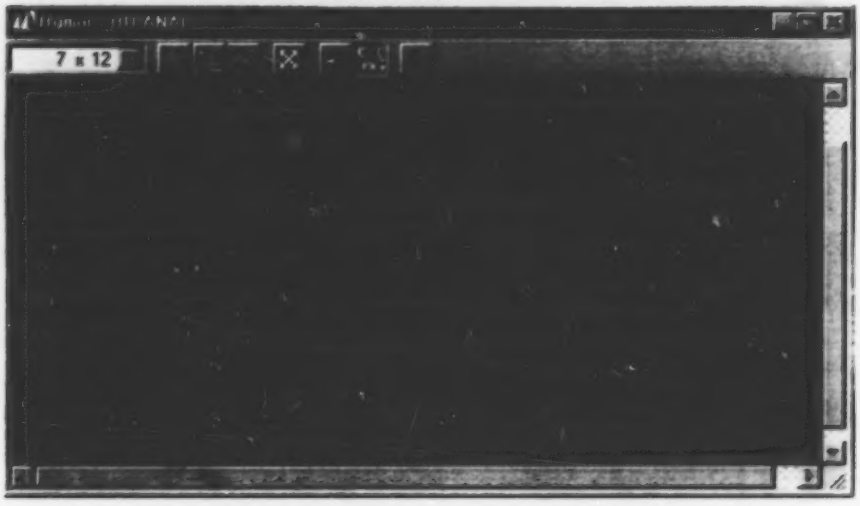

Figure 1. Hungarian morphological analysis with Humor

Humor's Hungarian version – the largest and most precise implementation – contains nearly 100 000 stems which cover all (approx. 70 000) lexemes of the *Concise Explanatory Dictionary of the Hungarian Language*. Suffix dictionaries contain all the inflectional suffixes and the productive derivational morphemes of present-day Hungarian. With the help of these dictionaries, **Humor** is able to analyze and/or generate several billions(!) of different well-formed Hungarian word-forms.

2. HelyesLem, the Lemmatizer

Humor has been rigorously tested both by linguists and "real" end-users of word-processing tools, namely. **Humor**-based linguistic modules have been licensed by *Microsoft, Lotus, Inso,* and other software developers. Another sort of testing of MorphoLogic's lemmatizer, **HelyesLem** (Figure 2), has also been used in every-day work since 1991 by both lexicographers and other researchers of the Research Institute of Linguistics of the Hungarian Academy of Sciences.

Figure 2. Stemming sample run of HelyesLem

The lemmatizer shares some of the extra features of Helyes-e?, the speller derived from **Humor**, because lexicographers need a fault-tolerant lemmatizer that is able to overcome simple orthographic errors and frequent mis-typings. Thus, it is able to treat "nonstandard" text e.g., from 19th century, when the orthographic system was not standardized as nowadays. Word-forms "auto-corrected" into the standard orthography can be properly analyzed (Figure 3).

3. Disambiguator, tagger, and parser in one tool: HumorESK

The basic strategy of **Humor** is inherently suited to parallel execution. Search in the main dictionary, secondary dictionaries, and affix dictionaries can occur simultaneously. What is more, in the near future, it is going to be

extended by a disambiguator based on the same strategy. This is a new parallel processing method of various levels (higher than morphology) called **HumorESK**, that is, **Humor** Enhanced with Syntactic Knowledge (Figure 4). Both **Humor** and **HumorESK** have a very simple and clear strategy based on surface-only analyses: no transformations are used; all the complexity of the systems are hidden in the graphs describing morpho-syntactic behavior.

Figure 3. Fault-tolerance of HelyesLem

Figure 4. Parsing Hungarian with HumorESK

4. Languages

The **Humor** description has been made for various natural languages. Both the size of the dictionaries and the depth of the description vary, but on the basis of the experiences with them, further languages can also be described. The list of the currently existing natural languages is shown by Table 1.

East-European languages:	Ready:	Hungarian Polish Romanian
	Under development:	Bulgarian
	In preparation:	Ukrainian Czech
Non-East-European languages	Ready:	English
	Under development:	German French
	Demo versions:	Italian Latin Ancient Greek

Table 1. Languages supported by Humor

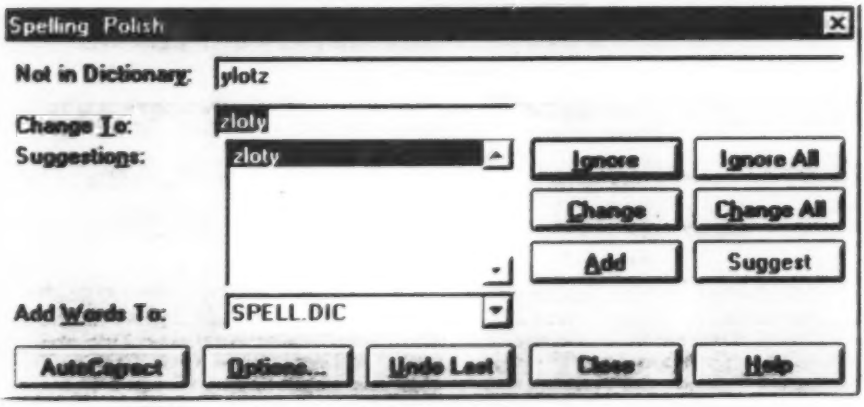

Figure 5. Polish speller based on Humor

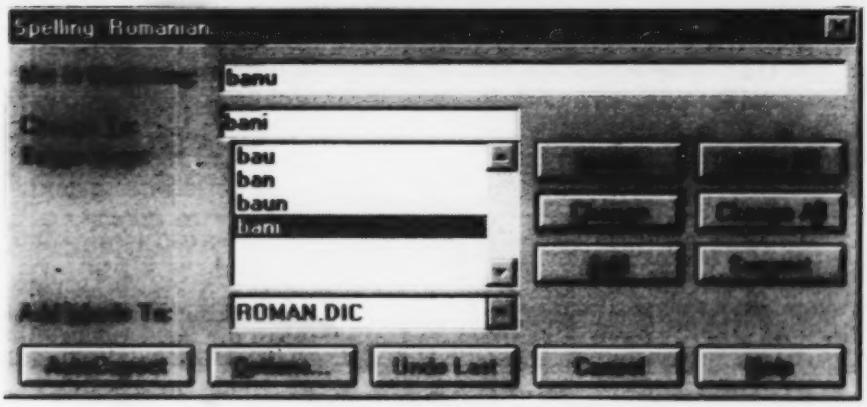

Figure 6. Romanian speller based on ***Humor***

5. Products on the Market

There are several commercially available **Humor** subsystems for different purposes: lemmatizers, hyphenators, spelling checkers/correctors and grammar checkers. They (called HelyesLem, Helyesel, Helyes-e? and Helyesebb respectively) have been built into several wordprocessing and full-text retrieval systems. Among others, Microsoft and Lotus licensed them for all of their localized Hungarian products.

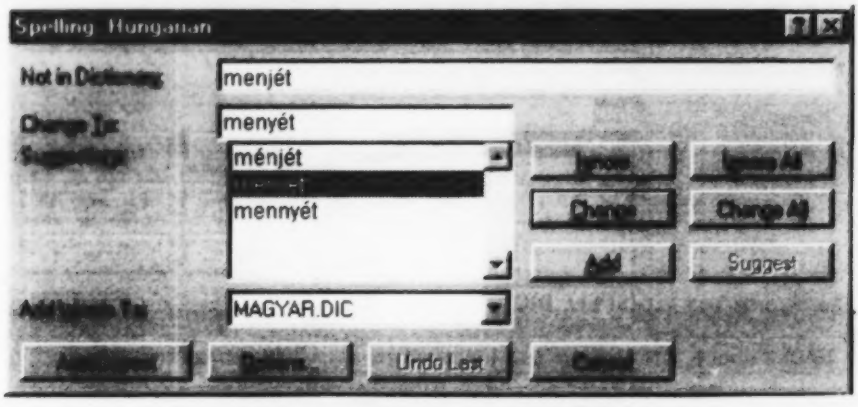

Figure 7. Hungarian speller based on Humor

BEST COPY AVAILABLE

Software products supported by **Humor**-based tools contain almost all of the Windows- and Macintosh-based word-processors (Word for Windows, AmiPro/WordPro, WordPerfect, etc.), DTP systems (Quark Xpress, Adobe PageMaker, Corel Ventura), and other applications (Excel, PowerPoint, Access, etc.) as well. Stand-alone versions work with pure text files and files in some formatting languages like RTF, PDF, SGML etc. (Figure 8)

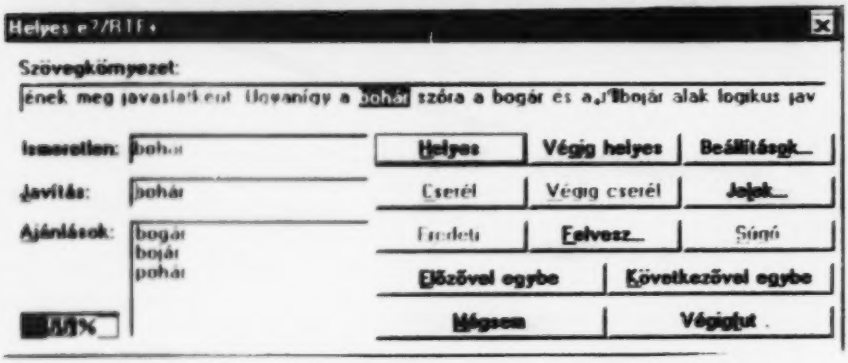

Figure 8. Application-independent (RTF) speller for Hungarian based on Humor

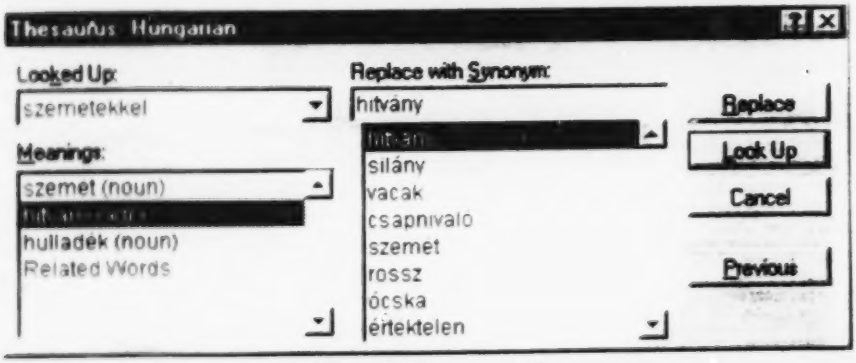

Figure 9. Hungarian thesaurus with stemming

Besides the above application there are two new tools based on the same strategy, the inflectional thesaurus called Helyette (Prószéky & Tihanyi 1992) and the series of intelligent bi-lingual dictionaries called MoBiDic. Both are dictionaries with morphological knowledge: Helyette is monolingual, while – as its name, MorphoLogic Bi-lingual Dictionary, suggests – bi-lingual. Having analyzed the input word the both systems look for the found stem in

the main dictionary. The inflectional thesaurus stores the information encoded in the analyzed affixes and adds to the synonym word chosen by the user. The synthesis module of **Humor** starts to work now, and provides the user with the adequate inflected form of the word in question. This procedure has a great importance in case of highly inflectional languages.

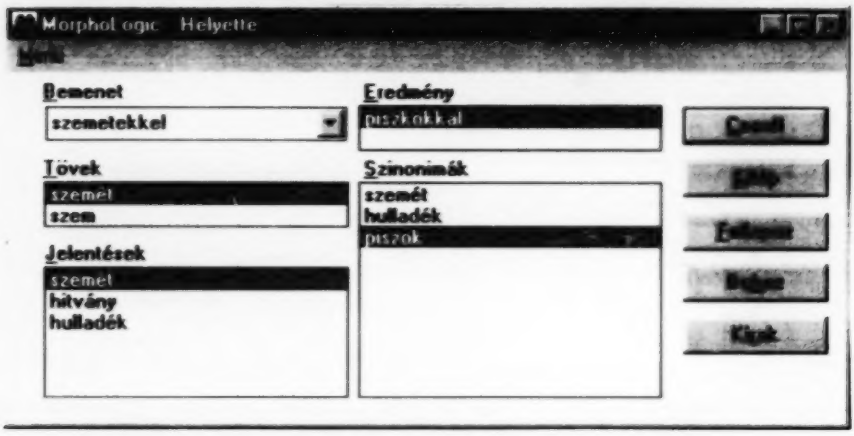

Figure 10. Hungarian thesaurus with morphological analysis and generation

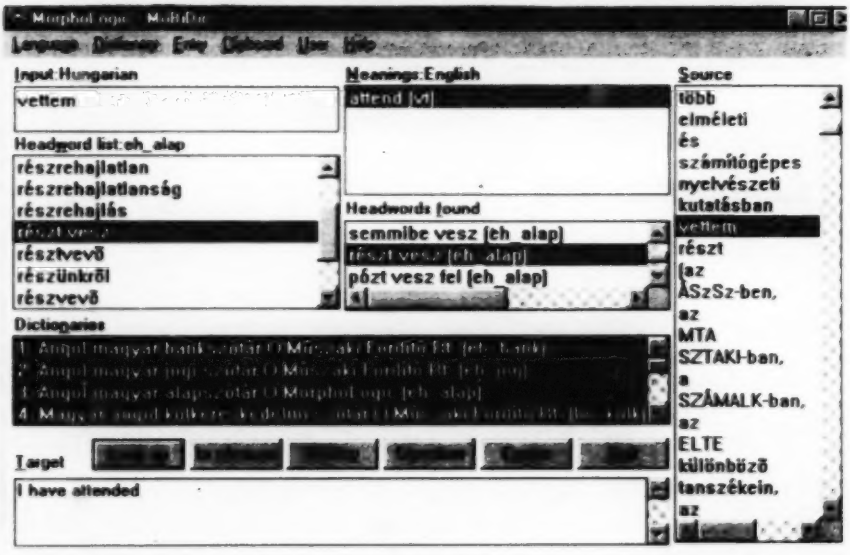

Figure 11. MoBiDic, the bi-lingual dictionary system with morphological modules

The synonym system of **Helyette** contains 40.000 headwords. The first version of the inflectional thesaurus **Helyette** needs 1.6 Mbytes disk space and runs under MS-Windows. The size of the **MoBiDic** packages vary depending on the applied terminological collection. E.g., the Hungarian–English Business Dictionary needs 1.8 Mbytes space.

MorphoLogic systems supporting OCR, hand-writing and speech recognition are under development. A sample run is shown on Figure 12. with the following non-segmented input:

kéremkészítsékelőjegyeiketésbérleteiketmertazállomásterületénellenőrökdolgoznak

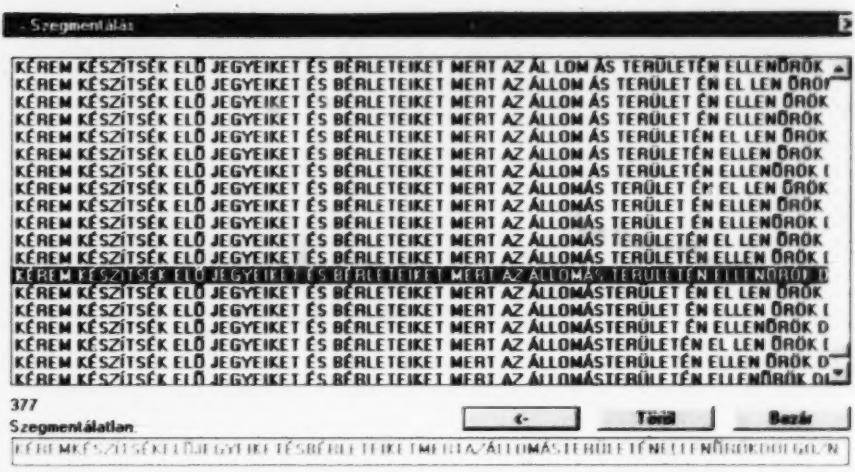

Figure 12. Segmentation of contiguous input based on ***Humor***

6. Technical details

The entire software package is written in standard C using C++ like objects. Portability to any platforms is guaranteed. The system runs on any platform where a C compiler can be found. MorphoLogic modules have existing implementations for DOS, Windows 3.1, Windows for Workgroups 3.11, Windows 95, Windows NT, OS/2, Macintosh, PowerMac, and Unix.

BEST COPY AVAILABLE

References

Prószéky, G. 1994. Industrial Applications of Unification Morphology. Proceedings of the 4th Conference on Applied Natural Language Processing, Stuttgart, Germany: 157–159.

Prószéky, G. and L. Tihanyi. 1992. A Fast Morphological Analyzer for Lemmatizing Corpora of Agglutinative Languages. F. Kiefer, G. Kiss & J. Pajzs (eds). Papers in Computational Lexicography – COMPLEX '92. Linguistics Institute, Budapest, Hungary: 265–278.

Prószéky, G. and L. Tihanyi, L. 1991. Helyette: An Inflectional Thesaurus for Agglutinative Languages. Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics: 473.

Prószéky, G. L. Tihanyi and M. Pál. 1994. **Humor**-based Applications. Proceedings of COLING-94, Kyoto, Japan: 1241–1244.